Chapter 15

Collusion–Free Privacy Preserving Data Mining

T. Purusothaman
Government College of Technology, India

M. Rajalakshmi
Coimbatore Institute of Technology, India

S. Pratheeba
Indian Institute of Science, India

ABSTRACT

Distributed association rule mining is an integral part of data mining that extracts useful information hidden in distributed data sources. As local frequent itemsets are globalized from data sources, sensitive information about individual data sources needs high protection. Different privacy preserving data mining approaches for distributed environment have been proposed but in the existing approaches, collusion among the participating sites reveal sensitive information about the other sites. In this paper, the authors propose a collusion-free algorithm for mining global frequent itemsets in a distributed environment with minimal communication among sites. This algorithm uses the techniques of splitting and sanitizing the itemsets and communicates to random sites in two different phases, thus making it difficult for the colluders to retrieve sensitive information. Results show that the consequence of collusion is reduced to a greater extent without affecting mining performance and confirms optimal communication among sites.

INTRODUCTION

Major technological developments and innovations in the field of information technology have made it easy for organizations to store a huge amount of data within its affordable limit. Data mining techniques come in handy to extract useful information for strategic decision making from voluminous data which is either centralized or distributed (Agrawal & Srikant, 1994; Han & Kamber, 2001).

The term data mining refers to extracting or mining knowledge from a massive amount of data. Data mining functionalities like association rule mining, cluster analysis, classification, prediction etc. specify the different kinds of patterns mined.
Association Rule Mining (ARM) finds interesting association or correlation among a large set of data items. Finding association rules among huge amount of business transactions can help in making many business decisions such as catalog design, cross marketing, etc. A best example of ARM is market basket analysis. This is the process of analyzing the customer buying habits from the association between the different items which is available in the shopping baskets. This analysis can help retailers to develop marketing strategies. ARM involves two stages

i) Finding frequent itemsets
ii) Generating strong association rules

**Association Rule Mining: Basic Concepts**

Let \( I = \{i_1, i_2, \ldots, i_m\} \) be a set of \( m \) distinct items. Let \( D \) denote a database of transactions where each transaction \( T \) is a set of items such that \( T \subseteq I \). Each transaction has a unique identifier, called \( TID \). A set of item is referred to as an *itemset*. An itemset that contains \( k \) items is a \( k \)-*itemset*. *Support* of an itemset is defined as the ratio of the number of occurrences of the itemset in the data source to the total number of transactions in the data source. Support shows the frequency of occurrence of an itemset. The itemset \( X \) is said to have a support \( s \) if \( s\% \) of transactions contain \( X \). The support of an association rule \( X \rightarrow Y \) is given by

\[
\text{Support} = \frac{\text{Number of transactions containing } X \cup Y}{\text{Total number of Transactions containing } X}
\]

An itemset is said to be *frequent* when the number of occurrences of that particular itemset in the database is larger than a user-specified minimum support. *Confidence* shows the strength of the relation. The confidence of an association rule is given by,

\[
\text{Confidence} = \frac{\text{Number of transactions containing } X \cup Y}{\text{Total number of Transactions containing } X}
\]

An association rule is said to be *strong* when its confidence is larger than a user-specified minimum confidence. Association rules with support and confidence above the minimum support and minimum confidence alone are mined. Many algorithms have been proposed for frequent itemsets generation. They are Apriori, Pincer search, Frequent pattern tree, etc. (Agrawal & Srikant, 1994; Lin & Kedem, 2002; Han, Pei, Yin & Mao, 2004).

**Distributed Data Mining**

In the present situation, information is the key factor which drives and decides the success of any organization and it is essential to share information pertaining to an individual data source for mutual benefit. Thus, Distributed Data Mining (DDM) is considered as the right solution for many applications, as it reduces some practical problems like voluminous data transfers, massive storage unit requirement, security issues etc. Distributed Association Rule Mining (DARM) is a sub-area of DDM. DARM is used to find global frequent itemsets from different data sources distributed among several sites and interconnected using a communication network. The DARM scenario is illustrated in Figure 1.

In DARM, the local frequent itemsets for the given minimum support are generated at the individual sites by using data mining algorithms like Apriori, FP Growth tree, etc. (Agrawal et al., 1994; Han et al., 2001). Then, global frequent itemsets are generated by combining local frequent itemsets of all the participating sites with the help of distributed data mining algorithm (Cheung, Ng, Fu, & Fu, 1996; Ashrafi, Taniar, & Smith, 2004). The strong rules generated by distributed association rule mining algorithms satisfy both minimum global support and confidence threshold.